2

AURORAL PARTICLES

David S. Evans
Space Environment Laboratory
National Oceanic and Atmospheric Administration
Boulder, Colorado 80303
and
Lockheed Palo Alto Research Laboratory
Palo Alto, California 94304

1. INTRODUCTION

While aurora have been a fashionable subject for study since the advent of the space age, a great deal of insight as to the origin of aurora had been gained in the years between the turn of the century and the Second World War, especially by Scandinavian physicists. It is illuminating to begin by reviewing their work as an illustration of the progress that can be made toward understanding a difficult subject through careful analysis and interpretation of observations. Although their methodology was based on the most advanced technology of the time, it would be considered entirely inadequate by today's standards. The background for this introduction may be found in books by Harang [1951], Störmer [1955], Chamberlain [1961], and Eather [1980].

Störmer's program of auroral photography, begun shortly after 1900, was directed toward defining the geometric shapes and locations of auroral forms. The best remembered of his experimental results concerned the distribution in altitudes of auroral features as derived from triangulation of photographic images taken of the same form against a star background from widely separated

points on the ground. As a general rule, the feature most easily identified as being common to the same form was the lower altitude border, although triangulations on other features, such as the high altitude boundary of extended rays, were also performed. The results showed that the lower altitude border of discrete auroral forms tended to have a distribution which maximized at about 105 km, with individual measurements extending down to 65 km and upwards to many hundreds of kilometers. During the same era, Vegard, also of Norway, made observations of the vertical extent of auroral luminosity and found these dimensions averaged some tens of kilometers but were hundreds of kilometers high in exceptional cases. Vegard also measured the vertical distribution of luminosity along an auroral form and showed that normally there was a maximum at an altitude some 10 km above the lower border.

Relatively little seems to have been done early in this century to quantify the horizontal dimensions of auroral forms. It was, however, clearly appreciated that the dimensions were usually very long in one direction (hundreds of kilometers east-west) and very thin in the other (tens of kilometers down to a small fraction of a kilometer for the structure in an individual auroral ray).

The period between 1890 and 1910 also saw the discovery of subatomic charged particles and radioactivity. Laboratory experiments had been conducted to study the properties of cathode rays (electrons), their penetrating power through materials, and to demonstrate how the rays interacted with a magnetic field. Similar experiments studied the nature of the particles emitted from radioactive materials, named alpha and beta rays, as well as the properties of fast ionized hydrogen atoms. Much work was also done with glow discharge tubes where fast subatomic particles interacted with gases resulting in the emission of light. Before 1900, Birkeland suggested that the aurora was produced by such subatomic particles transiting from the Sun to the Earth. Not only would this explain the emission of the light by processes analogous with those occurring in a gas discharge tube, but would also account for the observation that auroral rays aligned themselves along the geomagnetic field [a fact first noted by Wilcke in 1977] because these charged particles would be guided into the atmosphere along the nearly vertical magnetic field. The proposal

that the aurora was due to the impact of subatomic charged particles upon the atmosphere was generally accepted within a short time.

The altitude measurements of the aurora could be combined with the laboratory observations of the stopping power of subatomic particles passing through gases and with rudimentary models of atmospheric densities at high altitudes (based upon hydrostatic equilibrium arguments) to obtain estimates of the energies required for the incident particles to penetrate to the observed altitudes. Calculations done by Lenard, Vegard, and others before 1920 suggested that if the responsible particles were cathode rays, then the energies required to penetrate to 100 km would be on the order of 10 keV. If the particles were protons, the energies would be on the order of 200 keV. If alpha particles were hypothesized, the energies would need to approach 1000 keV. However, it was generally believed at this time that the responsible particles were cathode rays. There were two reasons for this. The first involved the height-luminosity profile of the aurora. If positive particles were responsible, then the luminosity profile of the aurora should exhibit a dramatic brightening with decreasing altitude and a very well-defined lower border, features that replicate the energy loss characteristics of positive particles passing through gases. These features were generally not present in the auroral luminosity profile which spoke in favor of cathode rays as being responsible. The second reason centered around the horizontal dimensions of auroral structures which were often very small. It was argued that these dimensions should in some respect be related to the gyro-radius of the responsible particles as they moved downward in the geomagnetic field. It was often impossible to reconcile the small horizontal dimension, which suggested that if massive positive particles were involved, they were relatively low velocity, with the altitude of the aurora which would require fairly energetic positive particles. If cathode rays were responsible, there would be no problems of this sort.

For a variety of reasons, speculation about the ultimate origin of the cathode rays centered around the Sun. First of all, evidence pointed toward an extra-terrestrial source because auroras were almost exclusively phenomena occurring at high geomagnetic latitudes, and particles approaching the Earth from infinity would naturally be guided to high latitudes and excluded from low

latitudes by the geomagnetic field. Secondly, it had long been noted that the occurrence of auroras followed events on the Sun. The frequency of auroras duplicated the 11-year cycle in sunspot numbers and, equally important, unusually intense aurora and geomagnetic disturbances often followed solar flares by one or two days. Finally, correlations between the passage of sunspot groups past the central meridian of the Sun and aurora at the Earth showed that increases in auroral occurrence followed the sunspot passage by a day or so. By 1920, the picture emerged of the Sun ejecting streams of particles which transited to the Earth where they produced auroral displays. It was with this picture in mind that Störmer began his study of the orbits of charged particles moving from infinity into a dipole magnetic field. His objective was to demonstrate that particles originating from the Sun would impact the Earth's atmosphere at locations where the aurora was observed. His orbit tracing showed that particles (electrons or ions) of the energies believed necessary to penetrate to the proper depths in the atmosphere would impact at locations very close to the magnetic pole and not at the most frequent latitude for auroras which is considerably displaced from the pole. Stormer attempted to escape this problem by modifying the geomagnetic field through the addition of a toroidal current around the Earth (the ring current), but, even then, he could not obtain acceptable results. A second problem that existed with the picture had to do with the one-day delay between an event on the Sun and the subsequent aurora at the Earth. If this time were ascribed to the transit time for the particles responsible for the aurora, the particle velocities and energies were far too low to be identified with the supposed auroral particles; e.g., 10 eV electrons based upon transit time arguments against 10 000 eV electrons inferred from the altitudes of aurora and similar discrepancies, if the heavier positive ions were assumed.

By the start of the Second World War, it is fair to say that the following was thought to be "known" about the aurora. The aurora was caused by the impact of charged particles upon the atmosphere. These particles were probably electrons (cathode rays) and had energies on the order of 10 keV. The particles probably originated from the Sun and transited to the Earth in a time of about one day. It was appreciated, however, that this picture required the existence of an unknown near-earth process(es) which accelerated these particles to the energies required to produce the auroral form and that this same

process, or a second one, was responsible for establishing the geometric shapes and geographic location assumed by the aurora. Today, many of the justifications for current and future space plasma investigations are lineal descendants of these problems originally set down nearly 50 years ago.

2. PROGRESS FROM WORLD WAR II TO 1970

The first unambiguous evidence that energetic subatomic particles were participating in auroral displays was provided by Vegard and by Meinel in 1950. To the surprise of many, this evidence pointed toward energetic protons, rather than electrons, as the responsible particles. The observations were of line emissions from hydrogen atoms which had been Doppler-shifted to such an extent that the excited atoms must have been moving at thousands of km sec^{-1} (tens of keV energy) at the time of emission. Presumably, a proton entered the atmosphere with this magnitude energy, picked up an electron from within the atmosphere to become an excited neutral hydrogen atom, and had emitted the photon while still moving at high velocity. There was a period of time after these observations when it was thought by many that auroral light was produced primarily, if not exclusively, by proton bombardment of the atmosphere. This view did not last for long because the counter-arguments set down by the Scandinavian researchers before 1940 were far too compelling. Those original arguments, based upon the height-luminosity profiles and small scale dimensions often seen in auroral forms, had been bolstered by the fact that the intensity of the hydrogen emissions varied immensely compared to other auroral emission lines originating from normal atmospheric constituents such as atomic oxygen and molecular nitrogen. This could not be the case if proton bombardment were dominant.

The first direct measurements of the particles producing visible auroral displays were made by instruments on sounding rockets during the 1958 International Geophysical Year (IGY) program [Davis, Berg, and Meredith, 1960; McIlwain, 1961]. These rocket flights showed conclusively that the visible aurora was produced primarily by the precipitation into the atmosphere of electrons

having energies of the order of 10 keV. These rocket observations demonstrated that the energetic proton precipitation was responsible for only a small portion of the energy deposited into the atmosphere, and that the protons were incident over an area that extended well beyond that of the visible auroral forms.

The conclusions as to the nature of the particles causing the aurora that had been drawn by the Scandinavian physicists some 20 years earlier on the basis of indirect measurements were entirely vindicated by these rocket observations. McIlwain's work had particularly long lasting importance. The instrument on board the rocket used to measure the electron influxes was primitive by today's standards (the instrument was not capable of sensing electrons of energies less than 4 keV and only obtained rather crude energy flux versus electron energy distributions by means of a sweeping electromagnet) and the rocket performance was low, reaching only 120 km altitude. In spite of these limitations, McIlwain was able to combine the electron energy flux measurements with the altitude profile of the auroral luminosity obtained by a photometer on board the rocket to demonstrate that electrons of energies greater than 10 keV contributed less than 10% and electrons of energies less than 3 keV contributed less than 25% to the total particle energy flux incident upon the atmosphere. McIlwain characterized this electron energy distribution as being "near monoenergetic". He suggested further that this sort of energy distribution was not consistent with a "statistical type" acceleration mechanism, but that the "sharp high-energy cutoff" in the electron energy flux distribution was consistent with "acceleration processes involving electric fields".

Several years later and following the development of particle detectors better able to measure the fluxes of electrons over the energy range 100 eV to 10 keV, McIlwain's conclusion as to the near monoenergetic nature and origin of the auroral electron energy spectrum was fully accepted. Numerous measurements by rocket- and satellite-borne instruments have now shown that the type of electron energy spectrum first described by McIlwain is invariably observed above "discrete" auroral arcs. Figure 1, displaying the electron differential-directional number flux versus energy spectrum of the electrons observed over a bright auroral arc, is typical. Here, there is a peak in the distribution at about 10 keV with an extremely sharp decrease in intensity

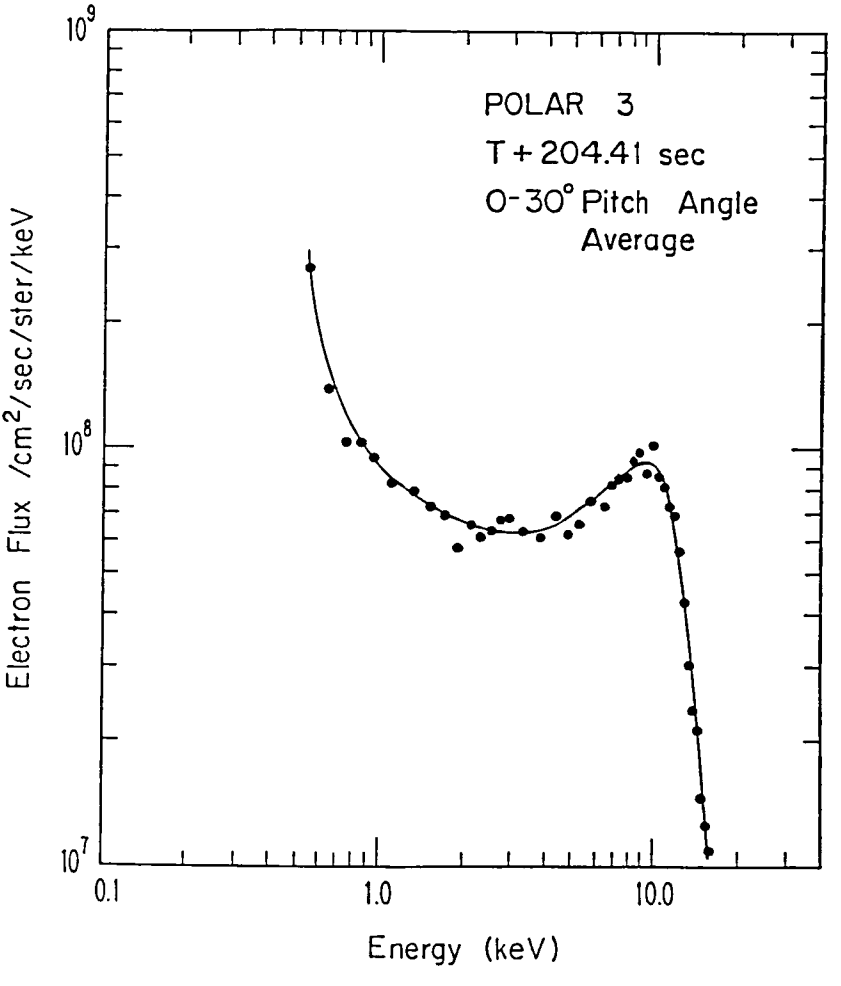

Figure 1. The electron number flux versus energy spectrum obtained above a discrete auroral arc. The spectral feature at near 10 keV is identified with the potential drop along the magnetic field line that accelerated the electrons. The lower energy electrons either originated from the atmosphere below the acceleration region or were accelerated electrons degraded in energy by some process. Positive ions played no role in the precipitation that produced this aurora.

at higher energies. While there are significant numbers of particles at lower energies, particularly below 1 keV, these particles contribute little to the incident energy flux. This form of energy spectrum may be fairly characterized as "near monoenergetic".

It should be pointed out that currently auroras are usually classified into two types, diffuse and discrete. As with most classification schemes, the separation is not totally unambiguous but does serve a useful purpose. Diffuse aurora tend to be widespread in spatial extent and uniform in intensity without well-defined boundaries to the light emission. It is this appearance which gives the name. Because of the absence of sharp boundaries, which would provide contrast against a dark sky background, this type of aurora is often difficult to recognize with the human eye. Moreover, the lack of easily identifiable features, such as rays or well-defined lower borders, make determinations of altitudes, locations, or height-luminosity profiles difficult to perform from the ground. For these reasons, relatively few rockets have been launched specifically to study such aurora, although satellite instrumentation has, on a regular basis, provided measurements of the particles producing such aurora.

Particle observations over diffuse aurora do show that energetic protons are participating in the bombardment, although it is rare that they carry more than 10% of the energy being deposited into the atmosphere, particularly when the total energy flux and auroral brightness are significant. The spectroscopic observations of Doppler-shifted hydrogen emissions made by Vegard, Meinel, and others were undoubtedly for cases of diffuse aurora. The energy spectra of both the electron precipitation responsible for the bulk of the energy deposition associated with the diffuse aurora and the lower intensity proton bombardment tend to resemble Maxwellian or "thermal-like" distributions without spectral features, such as the peak seen in the example in Figure 1. The mean energy of the electrons in the precipitation is ordinarily about 2 to 10 keV, while the mean energy of the protons is usually higher by a factor of 2 to 5. It is generally accepted that the electron and proton precipitation associated with the diffuse aurora originate from a reservoir of energized particles trapped in the geomagnetic field (the plasma sheet and outer radiation zones). These already energized particles are scattered in pitch angle by fluctuating electric fields and placed on trajectories leading into the atmosphere. The particle number densities and mean energies necessary on the part of the source plasma

population to supply the fluxes observed above diffuse aurora are in good agreement with the observed properties of the particle population in the plasma sheet. The contribution of protons to the total particle energy flux producing the diffuse aurora is also in quantitative agreement with this picture. For a source population composed of protons and electrons having equal densities and mean energies, and both species undergoing pitch angle scattering, the proton precipitation would account for about 2.5% of the total energy flux. If the proton temperatures were four times higher than the electron, the protons would contribute 5% under the same circumstances.

This explanation for the immediate origin of the particles producing the diffuse aurora does not address the question of the process whereby they acquired their energy. However, the Maxwellian-like nature of their energy spectrum suggests that the particles underwent collisional or randomizing processes as they were energized or afterwards. This would indicate an energization process which was either inherently statistical or, alternatively, a more ordered acceleration taking place in conjunction with statistical processes such as energy diffusion.

In contrast to the diffuse aurora, the discrete aurora has well-defined boundaries where the luminosity typically changes by a factor of 5 or more over a distance which is small compared to the overall dimensions of the form, a dimension which, in turn, is small compared to the dimensions of the diffuse aurora. The fact that discrete auroral forms exhibit such good contrast against a black sky allows easy identification of individual features for location and altitude determination. Most of the lower border altitude studies done by Störmer must have been of discrete forms. The excellent contrast of discrete auroras against a black sky combined with the fact that the amount of particle energy influx and auroral brightness is generally larger than for the diffuse aurora makes this type of aurora much easier to see and photograph. For these reasons, the bulk of ground-based studies have probably been performed on discrete aurora, and the majority of rockets have been launched over such aurora.

Measurements of the particles producing discrete aurora invariably show that precipitating electrons, of energies seldom exceeding 20 keV, are the dominant contributors to the energy deposition. The contribution due to proton

bombardment is virtually always less than 1% and often less than 0.1%. Moreover, the electron energy spectra essentially always show a peak or knee at some energy (Figure 1) or, on rare occasions, more than one energy. When displayed in terms of energy flux versus energy (Figure 1 displays number flux versus energy) the peak in the electron distribution is usually very dramatic. McIlwain's rocket flight, which encountered a near monoenergetic electron population, was over a discrete aurora. The nature of the electron energy spectrum together with the almost complete absence of protons contributing to the energy input has led most to conclude, as McIlwain did, that these electrons had acquired their energy by falling through an electrical potential difference. McIlwain did not speculate on the geometry of this potential difference or on the trajectory these electrons must have gone along between their arrival near the Earth (presumably from the Sun) and their deposition into the atmosphere. The search for answers to these questions have engendered considerable controversy in recent years.

3. THEORETICAL EXPLANATIONS FOR PARTICLE ENERGIZATION NEAR THE EARTH

If it is assumed that electrical potential differences exist within the magnetosphere surrounding the Earth, essentially three alternative geometries can be envisioned. The first is that the potential is distributed so that its gradient, the electric field, is everywhere perpendicular to the geomagnetic field. The second is a geometry in which some portion, if not all, of the available electrical potential is distributed parallel to the geomagnetic field. The third is a situation where the electric fields exist in a region where the magnetic field is zero, or nearly so.

Taylor and Hones [1965] explored the motion of charged particles in an electric and magnetic field geometry where the electric field was everywhere normal to the magnetic field. At that time, little was known of the nature of electric field that might surround the Earth and Taylor and Hones were compelled to develop a model based upon the electric fields that would be required to "drive" the currents known to flow in the ionosphere during times of magnetic

activity. The magnitude of these currents was estimated from the magnetic field perturbations that were observed from the ground and coupled with estimates of the electrical conductivity of the ionosphere to obtain the model of electric fields at ionospheric altitudes. These model electric fields were mapped upwards into the magnetosphere using a model magnetic field and assuming that the electric field always remained normal to the magnetic field. The creation of a model magnetic and electric field geometry for the entire magnetosphere was a major achievement for that time. Taylor and Hones then assumed the existence of a population of low energy particles at the boundary of their model magnetosphere (these particles having come from the Sun) and followed their trajectories in the model fields. In this particular geometry, the motion of individual low energy particles is a combination of a magnetically controlled drift, due to field line curvature and to gradients in the magnetic field, and an electrically controlled $\mathbf{E} \times \mathbf{B}$ drift. Given the gradients in the electric and magnetic fields and the scale size of the particle's gyroradius, the particle motion is adiabatic. The $\mathbf{E} \times \mathbf{B}$ drift alone is incapable of energizing particles because the drift path would be along an equipotential surface normal to $\mathbf{E}$. However, the superposition of the $\mathbf{E} \times \mathbf{B}$ drift and the magnetically controlled drift could carry the particles along a trajectory that has a component parallel to the electric field and result in the energization of the particle, effectively by moving through a potential difference. Particles of rather low solar wind energies can move through this geometry to a point where the electrostatic potential differs considerably from that at the entry point. The particle at this location will have its original energy plus that obtained by moving through the potential difference. If the latter exceeds the former by a significant amount, the particle energy spectrum at the final location will appear to be monoenergetic. If an individual particle is not precipitated into the atmosphere during its transit through the magnetosphere, its trajectory, being adiabatic, will return it back to the solar wind with its original energy. It should also be noted that the model leads to a separation between the trajectories followed by electrons and those followed by protons.

The model of Taylor and Hones explains, in a natural manner, both near monoenergetic particle spectra and the absence of energetic protons in the precipitation responsible for discrete auroras. However, the model does not easily account for the location and geometry of discrete auroral arcs. For a

particle source having a full range of incident pitch angles and energies at the magnetospheric boundary, the resultant trajectories do not form a line similar to an auroral arc but, rather, fill up much of the outer magnetosphere. The magnetosphere effectively acts as a crossed field particle analyzer with individual particles proceeding along trajectories to locations which are determined by their initial conditions (angle, velocity, mass, and charge). Taylor and Hones invoked a localized pitch angle scattering process, similar to that associated with the diffuse aurora, to precipitate the pre-energized particle population in the geometry appropriate to the discrete aurora. While this theory clearly describes one particle energization process which operates within the magnetosphere, it cannot easily account for those electrons producing discrete auroral arcs.

Speiser [1965, 1967] constructed a model in which an electric field was applied in a region of space, the magnetic tail, where the magnetic field was very small (a neutral sheet). Under these conditions, whether the electric field was perpendicular or parallel to the magnetic field was a moot point. Because in a near-zero magnetic field the dimensions of the particle's orbit would be large compared to the gradients in the electric and magnetic field, the motion of a particle would no longer be adiabatic. Using a tail-like magnetic field geometry and a dawn-to-dusk applied electric field, Speiser solved for the particle trajectories analytically. The results showed that a particle introduced into this geometry would undergo energization by moving parallel to the electric field while, at the same time, a north-south oscillatory motion between the tail lobes due to the particle's motion in the very weak neutral sheet magnetic field. Ultimately, the particle would either exit the system on the dawn (for electrons) or dusk (for positive ions) flanks, having been energized, or find itself in the tail lobes at a small pitch angle with respect to the magnetic field. Speiser showed that in the latter situation the particle, now energized, would follow a path along the magnetic field line toward the atmosphere. Speiser's model predicted near monoenergetic particle beams incident upon the atmosphere and that the electron and ion precipitation would be separated from one another. However, the model had difficulties in accounting for electrons of energies up to 10 keV without somewhat unrealistic assumptions about the magnetic field geometry. Essentially the electrons would be ejected from

the acceleration region very quickly and gain little energy from the electric field. Magnetic field configurations that would permit greater electron acceleration would result in those electrons entering the atmosphere at locations well poleward of where discrete aurora are usually observed. Finally, monoenergetic proton beams were also predicted but not observed even in proton-rich diffuse aurora.

Recently Lyons and Speiser [1982] expanded upon Speiser's original work and showed that if a plasma distribution having the number densities and temperatures of the plasma found in the "plasma mantle" were introduced into the neutral-sheet-electric-field acceleration geometry proposed by Speiser, the resultant ejected proton population calculated from the particle trajectories would have the intensities and energy distribution of those protons actually observed to be flowing on the outer edge of the plasma sheet. It is this population of positive ions that may be the major source of the plasma sheet population and, possibly, a direct source of auroral proton precipitation. The original Speiser model, perhaps in conjunction with additional particle energization by the Taylor and Hones adiabatic particle motion, may very well account for the energetic plasma population that forms the particle reservoir for the diffuse aurora precipitation where a significant admixture of protons is normally found. These same models, however, had difficulty in explaining the spatially structured, proton poor, and monoenergetic electron-rich character of the discrete aurora.

The third electric field acceleration geometry is one in which the electric field is directed parallel to the magnetic field. On the surface, this is a pleasing explanation. Assuming that the electric field is in the direction to accelerate electrons downward, electrons introduced across the high altitude boundary of the electric field will be energized and precipitated into the atmosphere in one direct process. The time taken for an individual electron to undergo the process is only seconds as opposed to a somewhat longer time for the original Speiser process and much longer for the Taylor and Hones adiabatic acceleration. In this picture, the geometry, small structure, and behavior of the discrete aurora are simply a manifestation of those magnetic field lines that possess

a parallel electric field, the total potential involved, and the time variations in that potential that may exist. The electron-rich nature of the discrete aurora and the monoenergetic spectrum are both direct consequences of the acceleration mechanism. By coupling the acceleration and precipitation process, the parallel electric field avoids the requirement invoked by Taylor and Hones for spatially structured pitch angle scattering processes to introduce corresponding spatial structure into the auroral precipitation. The particle acceleration by parallel electric fields may be located near the Earth instead of in the neutral sheet at great distances from the atmosphere as in the case of the Speiser model. This avoids the problem of accounting for small scale structure, introduced by the energization process, being preserved over long distances as the particles transit to the atmosphere. In spite of these seeming advantages, the idea that an electric field parallel to the magnetic field caused the energization and precipitation of those electrons responsible for discrete auroral arcs met with considerable resistance.

O'Brien [1970] summed up many of the arguments that a parallel electric field could not be the mechanism that energizes auroral particles. One point O'Brien stressed was that positive ions and electrons were observed to precipitate simultaneously. While this is usually the case for those particles producing the diffuse aurora (now interpreted as due to the loss of already energized particles from a reservoir, for example, the plasma sheet), the nearly monoenergetic electrons producing the discrete aurora seldom are accompanied by significant numbers of positive ions and do not seem subject to O'Brien's objection. A second, more telling, point made by O'Brien involved the presence of electrons in the precipitation which had energies lower than the magnitude of the accelerating potential difference that might be inferred from the location of the peak in the electron energy spectrum (e.g., those electrons of energies less than a few keV in Figure 1 where an accelerating voltage of 10 kV may be inferred). If the 10 keV electrons had fallen through a potential difference of that magnitude, then it seems that the lower energy electrons must have originated from within the region of parallel electric field and had acquired only a portion of the total available potential. However, as O'Brien pointed out, if this were the case, the large number fluxes of the low energy electrons

(see Figure 1) would require a powerful but unknown source of particles because the electrons locally available from within the region of parallel electric field would rapidly become exhausted.

Evans [1974] met this latter objection by pointing out that if there existed a parallel electric field which accelerated electrons of magnetospheric origin downward into the atmosphere, this same electric field would also reflect downward all secondary and backscattered electrons produced from the atmosphere by that precipitation. Effectively, the observed down-going electron population would be a combination of magnetospheric electrons energized by the electric field and electrons originating from the atmosphere that had been reflected back downward by that field. Numerical models were presented by Evans (Figure 2) which showed good agreement between an observed electron energy spectrum and that predicted using a backscatter-secondary model and a primary beam produced by accelerating a Maxwellian magnetospheric population through a fixed potential difference. Other comparisons between observation and model were not nearly so good. Generally in such instances, the predicted low energy electron fluxes were too low, particularly over energies between 20% and 80% of the accelerating potential (the spectrum in Figure 1 would likely be such an instance). A possible explanation for the excess of electrons over this energy range would be a process whereby the downgoing energized electrons produced beam-plasma instabilities in the ionosphere and the turbulent wave fields associated with this instability would diffuse electrons in energy, both promoting ionospheric electrons up in energy and degrading beam electrons down in energy [Evans, 1976]. In any case, the Earth's atmosphere and ionosphere represent a copious source of low energy electrons which would be confined below a parallel electric field, and so the existence of such electrons is not inconsistent with electron acceleration through a parallel electric field.

While O'Brien's objections, based upon observational considerations, to the existence of a parallel electric field which accelerated auroral electrons can be largely countered, there were also strong theoretical arguments against the very existence of electric fields parallel to a magnetic field, especially in the presence of a population of charged particles which were free to move under the influence of that electric field. The argument was that a static electric field

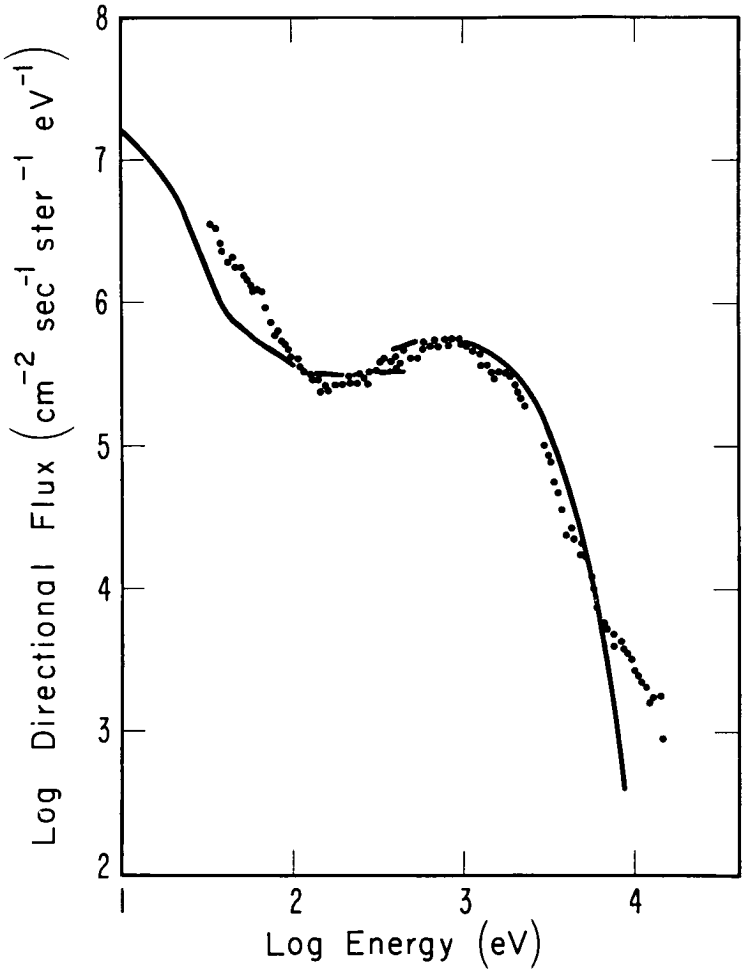

Figure 2. An example of a comparison between an observed auroral electron spectrum and one modeled by accelerating a Maxwellian distribution through a 500 volt field-aligned potential drop, computing the backscatter and secondary population created in the atmosphere by the precipitation, and reflecting that population from the potential barrier back downward. The good agreement overcame one of O'Brien's objections to the existence of parallel electric field acceleration.

parallel to a magnetic field also implied that a static charge distribution exists along the magnetic field line. However, the high mobility of electrons to move along the magnetic field (as opposed to across the field) under the influence of Coulomb forces would mean that such a charge distribution would be rapidly neutralized.

It was possible that an electrical current was flowing along the magnetic field line such that, as elementary charges moved out of a volume of space in a manner as to cancel the charge separation, new charges moved into that volume so as to maintain the charge distribution. In these circumstances, it was argued that the relationship between the current flowing along the magnetic field and the local electric field ought to be given by the Spitzer conductivity expression. Briefly, the Spitzer relationship between electric fields and currents had been derived in the following way. Consider a collection of ions and electrons, each with a number density, n, in the presence of an electric field. If there is a magnetic field also, it is assumed that the electric field is directed parallel to the magnetic field and the motion of charges in the direction normal to the magnetic field ignored. For simplicity, it is assumed that the ions are so massive in comparison to the electrons that their acceleration under the influence of electric forces is negligible and all current flow is due to the motion of electrons. The electrons are visualized as undergoing an acceleration due to the electric field but also a stopping or retarding force due to their occasional collisions with the massive slowly moving ions. This deceleration is expressed in this derivation as proportional to the electron's velocity $\vec{v}$ and a collision frequency, ν. Given this picture, the force balance equation may be written as

$$m \frac{d\vec{v}}{dt} = q\vec{E} - m\nu\vec{v} \tag{1}$$

If a steady state is to be established, then the average acceleration $\frac{dv}{dt}$ can be set to zero and a steady drift velocity on the part of the electrons with respect to the ions would be given by:

$$\vec{v} = \frac{q}{m\nu} \vec{E} \tag{2}$$

The electrical current, $\vec{J}$, produced by electrons of density, n, moving at this velocity is:

$$\vec{J} = nq\vec{v} \tag{3}$$

From equations (2) and (3) one obtains the Ohm's law relation between $\vec{J}$ and $\vec{E}$:

$$\vec{J} = \frac{nq^2}{mv}\vec{E}$$
$$\vec{J} = \sigma\vec{E} \tag{4}$$

Equation 4 was believed to govern the relationship between electric fields and currents parallel to the magnetic field. The conductivity relating the two would be determined by the number densities of charged particles and the frequency of collisions by these particles as they moved. The conductivity would be very large at high altitude along the magnetic field line because moving charged particles would suffer collisions only infrequently in low density medium. The frequency of collisions would be greater in the denser medium at lower, ionospheric level, altitudes; but even here the electrical conductivity parallel to the magnetic field would be very much larger than conductivity perpendicular to the magnetic field—charged particles being confined to move along the magnetic field and inhibited from moving normal to that field. The very large electric conductivities parallel to the magnetic field suggested that no significant portion of any available electrical potential could appear parallel to the magnetic field, but virtually all such potential must appear transverse to the magnetic field. An alternative argument examined Equation 4 in the limit where the collision frequency approached zero. In this limit, both the conductivity and the field-aligned currents would grow unbounded if the parallel electric field remained non-zero. This result was regarded as unphysical and was thought to prove that any electric fields parallel to the magnetic field must remain very small or zero. A significant parallel electric field would exist only if the conductivity could be reduced by some process. Models which invoked scattering (collisions) of charged particles by interactions with turbulent electric fields (anomalous resistivity) were proposed to reduce this conductivity. The basic purpose of such a process was to inhibit the motion of

a charged particle along the magnetic field and so "support" an electric field without the field-aligned current growing unbounded. However, in order to account for the acceleration of auroral electrons (which was the purpose of proposing a parallel electric field in the first place), the concept of runaway electrons—electrons which did not interact with the turbulent fields but accelerated freely in the large scale parallel electric field—was invoked. The final picture appeared forced and unsatisfying.

There were several errors in the above analysis. One was that in order to set down the equation of motion (Equation 1), it was assumed that the motion of the charged particles in the electric field was collisionally dominated, i.e., that the energy gained by a particle from the electric field between collisions and the energy lost by the particle in a collision were both small compared to the thermal energy of the charged particle. The validity of the Ohm's law expression in Equation 4 was predicated on this assumption, and it was improper to allow the collision frequency to approach zero as this limit was completely out of the range of applicability of Equation 2. A second error was to assume Equation 3 was valid in the case where there were no collisions. In such a case, the current would not increase as the charged particles velocity increased but, rather, the number density, n, would decrease in exactly the same manner that the density of automobiles on a freeway decreases as their speed increases after escaping from a traffic jam. The current, $\vec{J}$, would remain the same and be governed not by the electric field but by the rate at which new charged particles were allowed to enter the system. Another very basic error in the formulation of these arguments was the presumption that the amount of current that flows at a given location was governed solely by the local electric field (or that the local electric field was determined by the current flow). In fact, the analysis does not explain why either the electric field or the currents have any particular values, or even why they should exist at a given location. The electrical currents flowing and the electric field existing at a given location are determined not only by properties at that location but also by the nature of the rest of the electrical circuit under consideration. For example, the charge exiting a given volume (the current) is determined in the steady state by the availability of charge in the adjacent volume and its ability to flow so as to maintain current continuity. The analysis presented above simply assumed that sufficient charges were available. If they were not, the

values of the current and electric fields (which are physically independent of one another) would simply change so that current continuity was once again established.

A far more acceptable approach to the entire problem of electric fields parallel to the geomagnetic field and any associated current systems has been put forward by Knight [1973], Lemaire and Scherer [1974], Fridman and Lemaire [1980], Chiu and Schultz [1978], Lyons [1980, 1981], and many others. This approach begins by examining the current that can flow along the magnetic field between the magnetosphere and ionosphere in terms of the ability of a particle population at one location to supply charge (a current) to another location. The charged particles responsible for a current flowing upwards from the ionosphere to the magnetosphere must either be positive ions from the ionosphere or electrons from the magnetosphere. A downward current from the magnetosphere to the ionosphere must be carried by ions originating from the magnetosphere together with electrons from the ionosphere. The ability of each of these plasma reservoirs to supply charged particles to the other depends first upon the number densities and temperatures of the available plasmas and second upon the ability of charged particles to transit from one location to the other without returning back along the same path.

As an illustrative example, consider the maximum charge flux (current) that can flow from one location to the other in the absence of any parallel electric fields. For the conventional upward current, the contribution of ions from the ionosphere (assuming a density of 1000 cm^{-3} and temperature of 1 eV at 1000 km) can be no more than about 0.7 μAmp m^{-2}, this maximum being the rate at which these ions can evaporate from the top of the ionosphere by virtue of their thermal motion. The contribution to this current from electrons originating from the magnetosphere can be no more than that flux given by electrons filling the loss cone and is about 1 μAmp m^{-2} (assuming a magnetospheric density of 1 cm^{-3} and temperature of 1000 eV), which gives a total maximum upward current that these two populations can supply to one another of about 2 μAmp m^{-2}. A similar analysis for the maximum downward current yields a value of about 30 μAmp m^{-2}. The order of magnitude difference between the maximum upward and downward charge fluxes (currents) that can flow between the magnetosphere and the ionosphere is due to the ionosphere's ability to supply an upward flux of electrons which

is much greater than the upward ion flux because of the low electron mass and high thermal speeds for a given temperature. Note that these estimates are for the maximum charge fluxes that can flow between the two regions in the absence of a parallel electric field, not the currents that actually do flow in any situation.

Knight [1973] and Lemaire and Scherer [1974] extended this sort of analysis to the situation where a magnetic field-aligned potential difference was assumed to exist at some altitude well above the ionosphere. Both the change in particle trajectories because of this assumed electrical potential difference and the new field-aligned current can be calculated. It is clear that there is no change to the maximum upward going charged particle fluxes because this value is determined by the "evaporation rate" of these particles from the ionosphere. The assumed parallel electric field may accelerate the ionospheric particles upward but cannot change the fluxes. The flux of particles from the magnetosphere to the ionosphere will be changed because the acceleration of particles downward will effectively widen the loss cone and more magnetospheric particles of the species determined by the direction of the parallel potential drop will reach the ionosphere. However, Knight and Lemaire and Scherer showed that this effect is not large. If the parallel potential drop were located just above the ionosphere, only magnetospheric particles already magnetically mirroring at low altitude would have their trajectories affected, and the charge flux would be little affected. If the potential were assumed to be at high altitude, well removed from the ionosphere, many magnetospheric particles would be affected, but the analysis showed there would be only a modest increase in the flux of particles actually reaching the ionosphere and, thus, the field-aligned current. Most of the magnetospheric particles, even with an acceleration downward, would still magnetically mirror above the ionosphere and would return back to the magnetosphere having been decelerated to their original energy by the field-aligned potential difference. The magnetic field-aligned currents that would flow between the magnetosphere and the ionosphere in the presence of a magnetic field-aligned electric field would not grow unbounded, as the arguments based upon conductivities in a collisionless plasma would have suggested, but would assume values which would be governed largely by the ability of particle populations outside the region of electric field to supply charge. This ability might be quite limited. The situation that would

exist is quite analogous to that of a thermionic diode where the currents that can flow between the cathode and anode are governed not only by the direction and magnitude of the electric field between those two surfaces but also by the ability of the cathode to make free charges available to flow.

Two other points should be noted about this picture. First, placing a parallel electric field in the ionosphere, where large numbers of charges are available to flow, will not enhance the current between the magnetosphere and ionosphere. This current is determined by the ability (and requirement) for charges to exercise a trajectory which carry them irreversibly from one region to another and this very low altitude electric field will not influence the particles in the magnetosphere nor increase the fluxes of ionospheric particles upwards which will still be given by the rate at which ionospheric particles can migrate into the electric field region ("evaporation rate"). Secondly, there is a clear asymmetry between the maximum downward current that can flow between the magnetosphere and ionosphere (magnetospheric ions transiting to the ionosphere and ionospheric electrons transiting to the magnetosphere) and the maximum upward current that could exist. In this respect, the system also mimics the characteristics of a thermionic diode, both in the unidirectional nature of the current flow and in the fact that placing an electric field along the wire leading to the cathode of the diode will not increase the current that can flow between cathode and anode.

Lyons [1980, 1981] made use of this analysis of the ability of currents to flow between the magnetosphere and ionosphere to develop a model which can account for the existence of an electrical potential difference along the magnetic field connecting these two regions, the acceleration of magnetospheric electrons, and the creation of discrete aurora. The model presumes that at high altitude in the magnetosphere there is an electric potential distribution imposed over a limited region of space and in a direction perpendicular to the magnetic field. This potential distribution represents a source of electromotive force (EMF) capable of providing a dissipative current which threads both the ionosphere and the source of EMF to the extent that currents can flow along the magnetic field connecting the two regions. The reasons for the existence of this potential distribution are not specified, although electric field

measurements, both in the ionosphere and in the magnetosphere, show such distributions must occur.

Lyons' model imposes the requirement for current continuity on a current system which flows through the source of EMF, along the field lines to the ionosphere, and closes by flowing horizontally in the ionosphere between those magnetic field lines carrying the upward and the downward currents. He demonstrates, given the nature of the ionospheric and magnetospheric charged particle reservoirs, that current continuity ordinarily cannot be established if one presumes that the magnetospheric potential distribution is mapped unaltered along the magnetic field lines between the magnetosphere and ionosphere (i.e., if there were no parallel potential difference). Essentially, if the magnetospheric electric field were mapped directly into the ionosphere, a large ionospheric current would result because of the immense number of charge carriers capable of moving horizontally at ionospheric altitudes. These large currents would be inconsistent with the field-aligned currents that could be carried by the available charged particles moving between the ionosphere and magnetosphere. An alternative distribution of the available magnetospheric potential around this current circuit involving field-aligned potential differences would be required. Indeed, Lyons' model shows that the major portion of the available potential must appear along the magnetic field line. This arises because even a large parallel potential difference will not produce a dramatic increase in field-aligned currents and so a major reduction in the potential difference across the ionosphere would be required to bring about current continuity.

It is natural in this picture for the parallel potential difference to appear on that leg of the current circuit which is required to carry an upward current. It is this leg that has the poorest current carrying capability. The sense of the parallel potential that would appear would be to accelerate electrons downward into the atmosphere which, of course, is exactly what is observed. It is satisfying that this picture explains why discrete auroral arcs are produced by downward accelerated electrons and seldom, if ever, by ions which had been accelerated downward by a parallel potential in the opposite sense. In the auroral current circuit, as in laboratory current circuits, potential differences arise in those regions where the current carrying capability is minimal.

Lyons' model says nothing about the distribution of the parallel potential differences (i.e., the parallel electric field), only the necessity for that potential difference and an estimate of its magnitude. The detailed potential distribution is a matter related to the microphysics which govern the motion of particles along the field line—including the effects of the parallel electric field. It is of interest to point out that magnetospheric electrons being accelerated downward toward the atmosphere and ionospheric ions accelerated upward produce, by virtue of the velocity changes on the part of these particles as they move, a space charge distribution along the magnetic field which is in the proper sense to be responsible for the potential distribution (viz. net negative space charge at high altitude and net positive space charge just above the ionosphere). This illustrates a point which seems little appreciated. In current carrying circuits, it is the charge carriers and the details of their motion that are responsible for distributing the electric fields that not only ensure current continuity but also locally govern the motion of the charge carriers themselves. The fact that the current carriers also play the role of the space charges responsible for the local electric fields which govern their own motion may appear paradoxical. However, the very existence of the dissipative current system requires a source of EMF, and the charge carriers are best viewed in terms of distributing this EMF around the circuit in this case rather than creating electric fields.

4. SUMMARY

The problems concerning the aurora posed prior to the war are now either solved in principle or have been restated in a more fundamental form. The Scandinavians thought the charged particles responsible for the aurora had come from the Sun. While strictly speaking this may not be entirely true (ionospheric ions accelerated upward by a parallel electric field may populate the magnetosphere and reappear as auroral particles; electron backscatter and secondaries from the atmosphere may undergo the same recycling), it is generally agreed that the energy required to create the aurora, and the various other dissipative processes associated with the aurora, comes from the Sun in the form of the kinetic energy of charged particles transiting the interplanetary medium. The pre-war hypothesis concerning the nature of the auroral particles and their energies has been fully confirmed, with the exception that

helium and oxygen ions (presumably of ionospheric origin) were identified as participating in the auroral particle precipitation in addition to the protons. The nature of the near-earth energization processes affecting auroral particles has been clarified. These processes involve electric fields, a fact which would not have come as a surprise to the pre-war physics community. Charged particle trajectories in various electric field geometries have been modeled. An electric field in a region of zero or very low magnetic field near the Earth is very effective in energizing particles and populating a reservoir with hot plasma but, perhaps, not so effective in setting these particles on trajectories which lead directly to the creation of aurora. An electric field everywhere perpendicular to the magnetic field also is effective in energizing plasma trapped in that magnetic field. One or the other or both of these near-earth electric field geometries seem quite capable of creating a population of energized plasma which, as the particles are precipitated into the atmosphere, would create the diffuse aurora.

It has also been shown that electric potential distributions imposed perpendicular to the magnetic field in the outer magnetosphere can lead to electric field distributions along a circuit path that threads through the ionosphere. The major portion of the available potential is along the magnetic field line linking these two regions. Moreover, the sense of this field-aligned potential difference develops preferentially to accelerate electrons from the magnetospheric reservoir of hot plasma downward into the atmosphere. This accounts for all the important characteristics of the discrete auroral display, particularly the monoenergetic nature of the electron energy spectrum and the relative lack of positive ion participation in the particle bombardment.

The physical problems have now moved from determining the nature and geometry of the electric fields, which accelerate charged particles near the Earth, to accounting for the existence of these electric fields as a natural consequence of the solar wind's interaction with the Earth. These explanations will undoubtedly center around such physical situations as the creation of charge separations, the exchange of particle kinetic energy and electromagnetic potential energy, and the character of electrical current systems in unbounded space.

It is my opinion that ultimately the reward in continuing the work in auroral and magnetospheric particle dynamics will be a deeper understanding of the

subtleties of classical electricity and magnetism as applied to situations not blessed with well-defined and invariant geometries. Many of the concepts currently held as valid may fail us in this problem, simply because those concepts were predicated on certain aspects of a physical situation, such as wires which predetermine current paths, that must be relaxed. We have already seen how the concept of conductivity misled us in the analysis of electric fields parallel to the magnetic field line in the presence of a collisionless plasma. The idea—that electrical charges moving around a circuit act not only as current carriers but also through their own motion as the agents responsible for distributing the electric field in the proper manner to ensure current continuity—has been clarified by consideration of auroral particle dynamics. Of course, this latter concept applies equally well in a laboratory circuit (as do all fundamental concepts in electricity and magnetism), although it is not emphasized because it seems unimportant to obtaining a solution to those problems. The unbounded space of the solar wind, magnetospheric, and ionospheric system is a problem in which all our familiar constraints must be relaxed. In this sense, it is a laboratory for the study of the interplay of mechanical and electrical processes in the purest of situations. As an understanding of this system is gained, it is inevitable that additional long believed concepts about the nature of electricity and magnetism in dynamical systems will need to be modified or discarded.

REFERENCES

Chamberlain, J. W., 1961, *Physics of the Aurora and Airglow* (New York: Academic Press).

Chiu, Y. T., and Schultz, M., 1978, *J. Geophys. Res.*, **83**, 629.

Davis, L. R., Berg, O. E., and Meredith, L. H., 1960, *Proc. COSPAR Space Science Symposium* (Amsterdam: North Holland Publishing Co.).

Eather, R. H., 1980, *The Majestic Lights* (Washington, DC: American Geophysical Union).

Evans, D. S., 1974, *J. Geophys. Res.*, **79**, 2853.

Evans, D. S., 1976, *Physics of Solar Planetary Environments*, ed. D. J. Williams (Washington, DC: American Geophysical Union).

Fridman, M., and Lemaire, J., 1980, *J. Geophys. Res.*, **85**, 664.

Harang, L., 1951, *The Aurorae* (New York: John Wiley & Sons).

Knight, L., 1973, *Planet. Space Sci.*, **21**, 741.

Lemaire, J., and Scherer, M., 1974, *Planet. Space Sci.*, **22**, 1485.

Lyons, L. R., 1980, *J. Geophys. Res.*, **85**, 17.

Lyons, L. R., 1981, *J. Geophys. Res.*, **86**, 1.

Lyons, L. R., and Speiser, T. W., 1982, *J. Geophys. Res.*, **87**, 2276.

McIlwain, C. E., 1961, *J. Geophys. Res.*, **66**, 2727.

O'Brien, B. J., 1970, *Planet. Space Sci.*, **20**, 1821.

Speiser, T. W., 1965, *J. Geophys. Res.*, **70**, 4219.

Speiser, T. W., 1967, *J. Geophys. Res.*, **72**, 3919.

Störmer, C., 1955, *The Polar Aurora* (London: Oxford University Press).

Taylor, H. E., and Hones, E. W., 1965, *J. Geophys. Res.*, **70**, 3605.